1-1-2020

Beyond Covariance: SICE and Kernel Based Visual Feature Representation

Jianjia Zhang

Lei Wang
University of Wollongong, leiw@uow.edu.au

Luping Zhou
University of Wollongong, lupingz@uow.edu.au

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Beyond Covariance: SICE and Kernel Based Visual Feature Representation

Abstract

© 2020, Springer Science+Business Media, LLC, part of Springer Nature. The past several years have witnessed increasing research interest on covariance-based feature representation. Originally proposed as a region descriptor, it has now been used as a general representation in various recognition tasks, demonstrating promising performance. However, covariance matrix has some inherent shortcomings such as singularity in the case of small sample, limited capability in modeling complicated feature relationship, and a single, fixed form of representation. To achieve better recognition performance, this paper argues that more capable and flexible symmetric positive definite (SPD)-matrix-based representation shall be explored, and this is attempted in this work by exploiting prior knowledge of data and nonlinear representation. Specifically, to better deal with the issues of small number of feature vectors and high feature dimensionality, we propose to exploit the structure sparsity of visual features and exemplify sparse inverse covariance estimate as a new feature representation. Furthermore, to effectively model complicated feature relationship, we propose to directly compute kernel matrix over feature dimensions, leading to a robust, flexible and open framework of SPD-matrix-based representation. Through theoretical analysis and experimental study, the proposed two representations well demonstrate their advantages over the covariance counterpart in skeletal human action recognition, image set classification and object classification tasks.

Disciplines

Engineering | Science and Technology Studies

Publication Details

Zhang, J., Wang, L., Zhou, L. & Li, W. (2020). Beyond Covariance: SICE and Kernel Based Visual Feature Representation. International Journal of Computer Vision,

Noname manuscript No.
(will be inserted by the editor)

Beyond Covariance: SICE and Kernel based Visual Feature Representation

Jianjia Zhang, Lei Wang†, Luping Zhou, and Wanqing Li

Received: date / Accepted: date

Abstract The past several years have witnessed increasing research interest on covariance-based feature representation. Originally proposed as a region descriptor, it has now been used as a general representation in various recognition tasks, demonstrating promising performance. However, covariance matrix has some inherent shortcomings such as singularity in the case of small sample, limited capability in modeling complicated feature relationship, and a single, fixed form of representation. To achieve better recognition performance, this paper argues that more capable and flexible SPD (Symmetric Positive Definite)-matrix-based representation shall be explored, and this is attempted in this work by exploiting prior knowledge of data and nonlinear representation. Specifically, to better deal with the issues of small number of feature vectors and high feature dimensionality, we propose to exploit the structure sparsity of visual features and exemplify sparse inverse covariance estimate (SICE) as a new feature representation. Furthermore, to effectively model complicated feature relationship, we propose to directly compute kernel matrix over feature dimensions, leading to a robust, flexible and open framework of SPD-matrix-based representation. Through theoretical analysis and experimental study, the proposed two representations well demonstrate their advantages over the covariance counterpart in skeletal human action recognition, image set classification and object classification tasks.

Keywords Covariance Matrix · Structure Sparsity · Sparse Inverse Covariance Estimate · Kernel Matrix · Visual Representation.

Jianjia Zhang
School of Computer Science, University of Technology Sydney, Sydney, NSW 2007, Australia.
E-mail: seuzjj@gmail.com

Lei Wang †corresponding author
School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia.
E-mail: leiw@uow.edu.au
Homepage: `https://sites.google.com/view/lei-hs-wang`

Luping Zhou
School of Electrical and Information Engineering, The University of Sydney, NSW 2006, Australia.
E-mail: luping.zhou@sydney.edu.au

Wanqing Li
School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia.
E-mail: wanqing@uow.edu.au

1 Introduction

As a fundamental mathematical concept, covariance matrix has long been used in all sorts of areas in computer vision. Based on a set of feature vectors, covariance matrix characterises the variance of each feature and the statistical relationship between different features. By applying this property to visual feature representation, a seminal work (Tuzel et al., 2006) published more than one decade ago proposes to compute covariance matrix as a region descriptor, based on a set of feature vectors (e.g., intensity of colour channels or Gabor filter response) extracted at each pixel in an image region. It is shown that such a descriptor can effectively characterise the visual content, conveniently fuse different features, and be efficiently calculated. Also, this descriptor is partially invariant to image rotation or scaling and is robust against outliers. Thanks to these merits, this covariance-based region descriptor has been applied to object detection, recognition and tracking (Tuzel et al., 2006; Porikli et al., 2006; Tuzel et al., 2008) and shown promising performance.

The past several years have seen an expansion of covariance representation[1] in vision applications. Instead of only acting as region descriptor, covariance matrix has now been used as a general feature representation and applied to various tasks, including face recognition (Pang et al., 2008b), action recognition (Yuan et al., 2009; Guo et al., 2010; Zunino et al., 2017), image set classification (Wang et al., 2012), shape retrieval (Tabia et al., 2014), image segmentation (Ionescu et al., 2015), and so on. Recently, covariance representation has been further integrated with deep learning techniques to attain better visual recognition performance (Koniusz et al., 2013; Ionescu et al., 2015; Lin et al., 2015; Gao et al., 2016; Feichtenhofer et al., 2016; Li et al., 2017). The expansion from region descriptor to general feature representation brings forth new issues to covariance representation. Firstly, many visual recognition tasks experience the problem of small sample size and high feature dimensionality. This problem makes the estimate of covariance matrix less reliable, affecting its effectiveness and precision as a representation. Furthermore, the scarcity of samples can even lead to matrix singularity. Secondly, covariance matrix only characterises the linear correlation of features. Although this might serve the purpose of simplicity or efficiency as a region descriptor, it is certainly inadequate from the perspective of a general visual feature representation. In addition, flexibility could be another important consideration for a general representation. For example, for visual recognition tasks with various domain knowledge, different similarity measures may be needed to evaluate the relationship between features. Nevertheless, covariance matrix, which only measures linear correlations, cannot be readily altered to meet this need and is therefore not sufficiently flexible in this sense.

The above three issues indicate that we need to move beyond covariance matrix and develop new SPD-matrix-based representations. The SPD property is desirable to be retained because SPD matrices reside on a Riemannian manifold, and a number of specific algorithms have been developed in the literature to process, compare or classify such matrices. Exploring new SPD-matrix-based representations could take advantage of the existing algorithms and achieve better pattern recognition performance. This is attempted in this paper via two perspectives. To address the first issue of unreliability and singularity caused by small sample, we propose to exploit the prior knowledge on high-dimensional visual features. The prior knowledge could be from the domain theory of a specific vision application, for example, the "structure sparsity" due to the tree-shaped configuration of human skeletons (Lehrmann et al., 2013), or from more general principles such as the "Bet on Sparsity" (Hastie et al., 2005) used to estimate the structure of high-dimensional data (It will become clear in Section 3.2). To exemplify this approach, we migrate from covariance matrix to its inverse, and develop sparse inverse covariance estimation (SICE) (Friedman et al., 2008) as a new SPD-matrix-based visual representation. This new representation brings the following advantages: i) it is more robust against the scarcity of samples and completely free of the issue of singularity; ii) by incorporating prior knowledge, it can more faithfully characterise the underlying relationship of high-dimensional visual features; iii) it leads to a significant improvement in recognition performance; and iv) as by-product, SICE has a clear advantage over covariance matrix in revealing the relationship of features for interpretation.

To address the (second and third) issues of effectively and flexibly modeling nonlinear relationship, we propose to utilise kernel matrix as a general feature representation. Instead of computing a kernel between a pair of *samples* (as in common kernel-based learning methods), we compute a kernel between a pair of *feature dimensions*. Conceptually, this implicitly maps each feature dimension onto a higher- (or infinite-) dimensional space and evaluates their linear relationship therein. As will be demonstrated, this gives an effective way to model more sophisticated relationship between features, and the covariance representation is just a special case corresponding to the use of a linear kernel. Furthermore, for a wide range of kernel functions, this kernel matrix is guaranteed to be nonsingular, regardless of how small the number of feature vectors is or how high the feature dimensions are. In addition, the availability of various kernel functions could provide great flexibility in modeling nonlinear feature relationship, and extracting different relationship is just a matter of changing the kernel function. Last but not least, this kernel-based representation incurs little computational overhead. It can be explicitly computed and has the same size as covariance representation, and therefore well maintains computational efficiency. Also, compared with SICE-based representation, it does not need to solve any optimisation problem and works well when feature dimensions are high, and attains even better recognition performance.

This paper is a significant extension of our previous work reported in an ICCV paper (Wang et al., 2015a). The extension is made in five aspects: i) Besides the idea of kernel-based representation in (Wang

[1] Throughout the paper, we use "covariance representation" as a short name of covariance-matrix-based representation.

et al., 2015a), this paper newly proposes to incorporate prior knowledge and utilise SICE as a visual feature representation. As will be demonstrated, it consistently achieves better recognition performance than covariance representation, indicating the importance of incorporating prior knowledge; ii) More discussions are conducted to gain insight into the proposed representations, including computational complexity and convergence analysis. Their properties are summarised in comparison with other competing methods; iii) Experimental study is significantly expanded. More state-of-the-art comparable methods published after our ICCV work are included and more modern data sets (SBU Kinect, NTU RGB+D and ILSVRC2012) and advanced deep learning features are used for evaluation. iv) Visualisation of the proposed SICE and kernel-based representations is provided in Section 6 and Appendices 7 to give an intuitive understanding of their differences and advantages with respect to covariance representation; v) Lastly, sections of introduction and related work are expanded and enhanced.

The contributions of this work are recapped as follows. (i) To the best of our knowledge, we are among the first ones to improve covariance-based visual representation from the perspective of incorporating prior knowledge, and utilise SICE as a new visual representation for this purpose. This method effectively mitigates the issue of unreliable covariance estimation in the presence of small number of feature vectors; ii) We apply the kernel trick in a different way to characterise nonlinear feature relationship, leading to a new SPD-matrix-based representation with several desirable properties; (iii) The proposed two representations achieve significant improvement over existing covariance representation and other comparable methods in multiple visual recognition tasks, i.e., skeletal human action recognition, image set classification and object classification, demonstrating the advantage and generality of the proposed two representations.

The remainder of this paper is organised as follows. Section 2 reviews the literature on the use of covariance matrix as visual representation, and points out the newly encountered issues. Sections 3 and 4 elaborate how to develop SICE and kernel matrix as a general feature representation and describe their properties and advantages. After that, Section 5 discusses on the computational issue, convergence, and the differences of our work from the literature. Experimental result is reported in Section 6 and conclusions are drawn in Section 7.

2 Related Work

In the field of computer vision and image analysis, the processing of covariance matrix data can at least be traced back to diffusion tensor imaging (Basser et al., 1994), where a 3×3 covariance matrix is used to describe the diffusion of water molecules at each voxel in the brain. Also, covariance matrix has been utilised as a representation of the brain connectivity network estimated from functional magnetic resonance imaging (Smith et al., 2011). In relation to feature representation of general visual data, covariance matrix was initially proposed as an image region descriptor (Tuzel et al., 2006). The idea behind it lies at that for a visual feature vector (say, formed by the color intensity values at each pixel), the variances and statistical correlation of feature components presented on *a given set* (say, an image region) could be used to characterise this set. Let $\boldsymbol{x}$ ($\boldsymbol{x} \in \mathbb{R}^d$) be a d-dimensional feature vector and $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ denote a collection of n feature vectors in a given set. A covariance matrix $\boldsymbol{C}$ is estimated by

$$\boldsymbol{C} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top, \tag{1}$$

where $\boldsymbol{\mu}$ is the sample mean of feature vectors. As a region descriptor, it has the following merits: being a natural manner to fuse different kinds of features; being robust against illumination change and outliers; allowing two regions of different sizes to be compared; having rotation invariance when rotation-independent features are used; and fast computation via integral images.

Since the work in (Tuzel et al., 2006), covariance representation has been applied to various applications in computer vision. We can roughly categorise these applications into two classes based on its timeline of development.

i) *As a region descriptor.* This dominates the early applications. The effectiveness of region covariance descriptor is firstly shown on object detection and texture classification (Tuzel et al., 2006). Following that, it has been applied to object tracking (Porikli et al., 2006), pedestrian detection (Tuzel et al., 2008), and face recognition (Pang et al., 2008b). In these applications, the visual features generally include the coordinates, intensity values, or filter responses extracted at or around a pixel, while an image region or the whole image is usually the set to represent. Upon covariance representation, the similarity of image regions is evaluated for matching, and image label is predicted for recognition.

Two characteristics can be observed from these applications: 1) fast computation of region covariance descriptors is highly essential, especially for the tasks like object detection and tracking; 2) the dimensions of the

covariance matrix are generally low (e.g., 5×5 or 8×8) since the features are usually about pixel-based information. Comparatively, the number of pixels in a region is large, providing an adequate number of feature vectors. As a result, the covariance matrix can usually be reliably estimated without the singularity issue since n is significantly greater than d.

ii) *As a general representation.* This has recently been seen in an increasing number of tasks. For human action recognition, a covariance representation called Cov3DJ is proposed to model a sequence of skeletal joint motions over time (Hussein et al., 2013). In image set classification (Wang et al., 2012), a feature vector is extracted from every image in an image set and its covariance matrix is computed to represent this set. A similar case is observed in gesture recognition (Cirujeda and Binefa, 2014), where the covariance matrix of frame-based features is used to represent a video sequence. Besides, a line of recent work (Ionescu et al., 2015; Lin et al., 2015; Gao et al., 2016; Feichtenhofer et al., 2016) utilises covariance matrix to pool deep neural network features in various visual tasks. Two new characteristics can be observed from these more recent applications.

1) The wider range of applications poses a challenge on covariance matrix with respect to its effectiveness as a general visual representation. Due to the diversity of applications, the requirement on modeling all sorts of complicated feature relationships becomes evident. As a result, new SPD-matrix-based representations with more expressive power are highly desired.

2) Features are not simply pixel-based information anymore and usually have higher dimensions. For example, in skeletal action recognition, the dimensions of feature vector can be as high as 120, while the total number of frames per action instance can be as low as 40. An even worse case is found in image set classification. The feature dimensions reach 400 (by reshaping a 20×20 thumbnail grey-level image to a vector), while there are only 41 images in a set (Wang et al., 2012). When the features extracted from deep neural networks are used, their dimensions could become even higher. In contrast to the previous case of region descriptor, the problem of small number of feature vectors and high feature dimensionality is now more frequently encountered, resulting in unreliable or singular covariance representation.

During the evolution from region descriptor to general visual representation, covariance representation has also been steadily improved. A significant progress is on the similarity evaluation of covariance representations. Powered by the theories of Riemannian manifold, this line of research has produced a number of more effective measures, including Log-Euclidean distance (Arsigny et al., 2006), Cholesky distance (Dryden et al., 2009), Power Euclidean distance (Dryden et al., 2009), Stein divergence (Sra, 2011), Log-Hilbert-Schmidt metric (Quang et al., 2014), and those learned from data in supervised manners (Wang et al., 2015b). Through these similarity measures, many vector-based algorithms, e.g., support vector machines (SVM) and sparse coding, have been extended for covariance representation to perform classification (Sra, 2011; Jayasumana et al., 2013) or regression (Harandi et al., 2012).

In terms of covariance representation itself, some approaches have been proposed to improve the quality of visual feature or image region upon which it is computed. For example, considering that Gabor features could extract more important information, they are used to replace the first- and second-order image gradients at each pixel to compute covariance matrix for face recognition (Pang et al., 2008b). To reduce the interference from the background to object tracking, pixels are weighted in the computation of covariance matrix (Wu et al., 2015), and the farther a pixel is from the centre of a region, the lower its weight is set. Similarly, in action recognition (Guo et al., 2010), to avoid background pixels, only the pixels whose temporal gradients are greater than a threshold (identified as "the pixel related to the movement") are used to compute covariance representation. Recently, the idea of covariance matrix based representation has been incorporated into deep networks as pooling methods (Koniusz et al., 2013; Ionescu et al., 2015; Lin et al., 2015; Feichtenhofer et al., 2016; Li et al., 2017; Wang et al., 2019b). Along this direction, the works in (Gao et al., 2016; Cui et al., 2017) further develop compact and kernel pooling methods, respectively, to derive approximate kernel mapping functions, which are embedded in deep networks as a pooling layer. Our proposed kernel-based representation method is different from these methods because our kernel function is applied to each pair of feature dimensions while (Gao et al., 2016; Cui et al., 2017) apply a kernel function to each pair of feature vectors.

Among the literature, the closely related works to ours are kernelised covariance methods from (Pang et al., 2008a), (Harandi et al., 2014b), (Cavazza et al., 2016), which attempt to model high-order statistics of features. Their idea is to map all the feature vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ in a given set to another feature space via a kernel function and calculate a covariance matrix therein. Although this approach is powerful in characterising nonlinear relationship of features, it results in a potentially *infinite-dimensional* covariance matrix, defined in the kernel-induced feature space, as representation. Because these covariance matrices cannot be explicitly

computed, (Pang et al., 2008a) and (Harandi et al., 2014b) derive special measures to evaluate the similarity between the implicit high-dimensional covariance matrices. The computational complexity of the measures for evaluating a pair of covariance matrices reaches $\mathcal{O}(n^3)$ due to the need of performing eigen-decomposition (Harandi et al., 2014b). This will become computationally expensive or even impractical when n increases, as demonstrated in our experiment. To maintain the computational efficiency, the work in (Cavazza et al., 2016) develops an explicit approximate mapping function to realise the kernel trick. Again, the essential difference between our kernel-based representation method and these methods is that our kernel function is applied to each pair of feature dimensions while they apply a kernel function to each pair of feature vectors.

We also notice that there is another line of work, e.g., those in (Koniusz et al., 2016; Cavazza et al., 2017a, 2019), extracting non-linear information from data via kernel linearisation or approximation. However, the kernel in these works plays a different role from those methods mentioned above and ours. Specifically, these works develop kernel linearisation or approximation methods to measure the similarity between a pair of samples. In contrast, those above methods from (Pang et al., 2008a), (Harandi et al., 2014b), (Cavazza et al., 2016) and our proposed kernel representation method are to represent an individual sample rather than measuring the similarity between two samples. In other words, those above methods and our method perform at the level of single sample representation while the methods in (Koniusz et al., 2016; Cavazza et al., 2017a, 2019) work at the level of pairwise sample similarity measure for classification.

Reviewing the literature shows that a key task in the recent development of covariance representation is to address the situation entangled by the presence of small number of feature vectors, higher feature dimensionality, and more complicated feature relationship to characterise. The following parts report our attempts in this regard.

3 Proposed SICE Representation

3.1 Motivation and basic idea

Covariance representation describes the underlying structure of visual features distributed over a set, by assuming a Gaussian model and using sample-based covariance estimate. As previously mentioned, due to the scarcity of feature vectors and high feature dimensionality, such a covariance estimate is not able to faithfully reflect the underlying data structure. As stated by a general principle on knowledge representation, information on a given learning task comes from both training examples and domain knowledge (Haykin, 1998). Particularly, when the former is inadequate, exploiting the latter becomes essential. With regard to the covariance representation, it means we shall make good use of prior knowledge available from specific tasks on the underlying structure of high-dimensional visual features.

3.2 SICE as feature representation (SICE-RP)

Sparsity (Huang et al., 2011) may be the most common prior knowledge on the structure of high-dimensional data, and it has been well applied to various vision tasks. In the terminology of probabilistic graphical model (Koller and Friedman, 2009), a distribution can be illustrated as a graph, with each node corresponding to a feature component and each edge indicating the presence of statistical dependence between the linked two nodes. In this case, structure sparsity means the sparsity of the graph, i.e., only a small number of edges exist. A typical example of such a situation is in skeletal human action recognition. Due to the tree-shaped kinematic configuration of the human body, only a small number of joints are *directly* linked. This induces structure sparsity when joint-based features are collectively used to model an action.

Imposing structure sparsity into covariance representation is not straightforward. Covariance matrix measures the pairwise correlation of features without discriminating direct and indirect correlation, so it is not sparse by nature in most cases. In order to model direct correlation among features, we have to resort to the inverse of covariance matrix. This is because the inverse covariance measures the partial (i.e., direct) correlation by factoring out the effects of other variables (Huang et al., 2010). This allows the sparsity prior to be conveniently imposed.

Let us assume that $\boldsymbol{x}$ follows a Gaussian model $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is covariance matrix and $\boldsymbol{\Sigma}^{-1}$ is its inverse. Each off-diagonal entry of $\boldsymbol{\Sigma}^{-1}$ measures the direct correlation between two features. It will be zero if features i and j are conditionally independent given all the remaining ones. The estimate of $\boldsymbol{\Sigma}^{-1}$, denoted by $\boldsymbol{S}$, has been well resolved in the literature by maximising a penalised log-likelihood of data, with a SPD constraint on $\boldsymbol{S}$ (Friedman et al., 2008; Huang et al., 2010). The optimal solution is called sparse inverse covariance estimate (SICE).

$$\boldsymbol{S}^* = \arg\max_{\boldsymbol{S}\succ 0} \quad \log\left[\det(\boldsymbol{S})\right] - \mathrm{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{S}) - \lambda\|\boldsymbol{S}\|_1, \qquad (2)$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample-based covariance estimate, while $\det(\cdot)$, $\mathrm{tr}(\cdot)$ and $\|\cdot\|_1$ denote the determinant, trace and the ℓ_1-norm of a matrix. Through the term of $||\boldsymbol{S}||_1$, structure sparsity is imposed on $\boldsymbol{S}$ to achieve more reliable and faithful estimation. The tradeoff between the degree of sparsity and the log-likelihood estimation is controlled by the regularisation parameter λ. Changing λ value will reveal the underlying structure at various sparsity levels, with a larger λ inducing a sparser $\boldsymbol{S}^*$. The maximisation problem in Eq.(2) is convex and can be effectively solved by the off-the-shelf packages such as GLASSO (Friedman et al., 2008).

With regard to covariance representation in this paper, we can readily apply Eq.(2) to a given feature set $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ to obtain the corresponding SICE. Using it to replace covariance matrix gives rise to a new SPD-matrix-based visual feature representation. Compared with its covariance counterpart, this new representation enjoys several desirable properties. Firstly, by incorporating prior knowledge, this new representation is more tightly coupled with the task and can therefore more faithfully characterise the relationship of high-dimensional visual features; Secondly, due to the constraint of $\boldsymbol{S} \succ 0$ required in SICE, $\boldsymbol{S}^*$ is guaranteed to be nonsingular, even when the sample-based covariance estimate $\hat{\boldsymbol{\Sigma}}$ is singular. It is more robust against the scarcity of feature vectors and completely free of the issue of singularity; Thirdly, this new representation leads to significantly improved recognition performance on a variety of benchmark data sets; Lastly, through SICE we can have better knowledge on the relationship, i.e., direct correlation, among features that cannot be obtained from covariance representation. This by-product is not only useful for interpretation, but could also provide important cues to design more compact representation in the future.

As an established technique, SICE has been used in a variety of applications such as modeling networks of gene expression (Banerjee et al., 2008), cell signaling (Friedman et al., 2008), brain connectivity (Huang et al., 2010), and so on. However, it has not been considered for SPD-matrix-based visual representation before. Our contribution in this part lies in showing the importance of exploiting prior knowledge in helping covariance representation battle with small number of feature vectors. This perspective is new in the literature of covariance-based visual representation. Also, with this perspective, we not only propose SICE as a new representation, but also clearly display why it is a better option than its covariance counterpart.

Finally, it is worth noting that a more general justification for applying structure sparsity to high-dimensional data comes from the "Bet on Sparsity" principle (Hastie et al., 2005). As indicated by this principle, when the direct correlation relationship among features is truly sparse, imposing such a prior will be appropriate and can better characterise the underlying data structure. When the direct correlation relationship is not truly sparse, we will not lose much either, because there is no way to recover the true structure in the case of small sample. More detailed explanation can be found in (Hastie et al., 2005).

4 Proposed Kernel based Representation

4.1 Motivation and basic idea

Although enjoying a number of merits, SICE shares one critical drawback with covariance matrix, that is, both of them are only able to capture the linear correlation between features. As a general visual representation, modeling only linear relationship significantly constrains its expressive power and in turn affects recognition performance. For example, for human actions, which are generated by a complex and time-varying non-linear dynamical system (Ali et al., 2007), it is certainly insufficient to only consider the linear correlation of skeleton joints when differentiating action patterns. Also, the features from different channels in the convolution layers of neural networks are not necessarily linearly correlated. Another drawback of SICE lies at that it has to be obtained by numerical optimisation. This makes it computationally less attractive, especially when a large-sized SICE is sought.

To address these issues, we further propose to use kernel matrix to replace covariance matrix as a general visual representation. To manifest our motivation and make the presentation self-contained, we show that covariance matrix only describes linear correlation of features as follows. Recall that $\boldsymbol{x}$ is a d-dimensional feature vector, and let $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$ denote a $d \times n$ data matrix. We define $\boldsymbol{f}_i^\top$ $(i = 1, \cdots, d)$ to be the ith row of this matrix, consisting of the n realisations of the ith feature. After centering, it can be written as $\bar{\boldsymbol{f}}_i = \boldsymbol{f}_i - \mu_i \mathbf{1}$, where μ_i is the mean of the ith feature while $\mathbf{1}$ is a column vector of "1"s. It is trivial to show that the (i, j)th entry of covariance matrix $\boldsymbol{C}$ (defined in Eq. (1)) is

$$c_{ij} = \left\langle \frac{\bar{\boldsymbol{f}}_i}{\sqrt{n-1}}, \frac{\bar{\boldsymbol{f}}_j}{\sqrt{n-1}} \right\rangle, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product. In other words, covariance matrix essentially implements a *linear* kernel function over scaled $\bar{\boldsymbol{f}}_i$ and $\bar{\boldsymbol{f}}_j$. This limits it in capturing linear correlation between features.

4.2 Kernel matrix as feature representation (Ker-RP)

As a natural remedy, we propose to substitute a *nonlinear* kernel function for the above linear one, and utilise the resulting kernel matrix, denoted by $\boldsymbol{M}$, as a general visual representation. The (i,j)th entry of $\boldsymbol{M}$ is

$$k_{ij} = \langle \phi(\boldsymbol{f}_i), \phi(\boldsymbol{f}_j) \rangle = \kappa(\boldsymbol{f}_i, \boldsymbol{f}_j), \tag{4}$$

where $\phi(\cdot)$ is an implicit nonlinear mapping and $\kappa(\cdot,\cdot)$ is the induced kernel function (Schölkopf et al., 2002; Vedaldi and Zisserman, 2012). It is easy to see that covariance matrix corresponds to a special case in which $\phi(\boldsymbol{f}_i) = (\boldsymbol{f}_i - \mu_i \mathbf{1})/\sqrt{n-1}$. **Note that** the mapping $\phi(\cdot)$ is applied to each *feature dimension* $\boldsymbol{f}_i$, rather than to each *feature vector* $\boldsymbol{x}_i$ as seen in the closely related works in (Pang et al., 2008a), (Harandi et al., 2014b) and (Cavazza et al., 2016) or each sample as in (Koniusz et al., 2016; Cavazza et al., 2017a, 2019). The size of kernel matrix $\boldsymbol{M}$ maintains to be $d \times d$, same as that of covariance matrix. The most significant advantage of using $\boldsymbol{M}$ lies at that we can have more flexibility to model the nonlinear relationship among features by utilising kernel functions.

In practice, applying the proposed kernel-based representation is easy. When we do not know (or are not particularly interested in) what kind of nonlinear relationship shall be modeled beforehand, any general-purpose kernel, such as the most commonly used Gaussian radial basis function (RBF) kernel $\kappa(\boldsymbol{f}_i, \boldsymbol{f}_j) = \exp(-\beta\|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2)$, can be employed. Also, when it becomes necessary, users are free to utilise special kernels in certain areas to serve their goals or can even directly learn kernel functions from given data. Such flexibility is clearly an advantage brought by using a kernel matrix as feature representation.

In addition to the above nice property of generability, RBF kernel representation is also a better choice than covariance representation in relation to the singularity issue. It is known that in the case of $d \geq n$, covariance matrix is bound to be singular. In contrast, the situation is more favourable for kernel matrix. A direct application of Micchelli's Theorem (1986) (Haykin, 1998) (page 264, Chapter 5) gives the following result for our case.

Theorem 1. *Let $\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_d$ be a set of different n-dimensional vectors. The matrix $\boldsymbol{M}_{d\times d}$ computed with a Gaussian RBF kernel $\kappa(\boldsymbol{f}_i, \boldsymbol{f}_j) = \exp(-\beta\|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2)$ is guaranteed to be nonsingular, no matter what values d and n are.*

This result indicates that we do not need to worry about the singularity issue at all, when a RBF kernel is used. According to Micchelli's Theorem, the inverse multiquadric kernel $\kappa(\boldsymbol{f}_i, \boldsymbol{f}_j) = (\|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2 + \beta^2)^{-\frac{1}{2}}$ also satisfies the above theorem. Actually, as pointed out in (Fasshauer, 2011), in addition to these two kernels, there is a large range of kernels holding this nice property, including radial kernels, translation invariant kernels, multi-scale kernels, power series kernels, and so on. The presence of these kernels provides users great freedom to choose the most appropriate one for a kernel representation, while staying free of singularity issue. In addition, when a new kernel is tried and the nonsingularity of kernel matrix is uncertain, we can always analyse it based on the definition of positive definiteness and/or append a regulariser to this matrix as a preemptive measure.

Compared with SICE representation, the nonsingularity of kernel-based representation is naturally guaranteed through the use of (certain) kernels, instead of the SPD-constrained optimisation. This is not only easy to implement but also incurs no extra computation. Especially, when the feature dimensions d is high, kernel-based representation could still function well while SICE representation may become hard to obtain through optimisation. Also, kernel-based representation can generally lead to better classification performance than the SICE representation, as will be demonstrated shortly. On the other hand, SICE representation achieves the improvement over covariance representation by still rooting in linear techniques. More importantly, it actively exploits prior knowledge (e.g., structure sparsity) of a given task, which we believe is an important research direction but has not been well reflected in the kernel-based representation. In addition, the partial correlation characterised by the SICE representation can better reveal the essential relationship of features. Therefore, SICE representation has different properties from the kernel-based one, and it could be regarded as a capable linear alternative to the latter.

5 Discussion

5.1 Computational issues

The computation on SICE representation is discussed as follows. Given a $d \times d$ covariance matrix $\hat{\boldsymbol{\Sigma}}$ estimated from n feature vectors, the optimisation in Eq. (2) is proved to be convex and guaranteed to converge even in the case of $n < d$. The optimal SICE matrix can be effectively obtained by the off-the-shelf package like GLASSO (Friedman et al., 2008) in $\mathcal{O}(d^3)$. As shown in (Friedman et al., 2008), it takes only 0.497 CPU second to obtain a 200×200 SICE matrix on an Intel Xeon 2.80 GHz processor. Therefore, the proposed SICE representation will not incur significant computational load, when d is no more than 200. Also, note

that the computational cost of solving the optimisation in Eq. (2) is independent of the feature number n since it is based on a pre-calculated $\hat{\boldsymbol{\Sigma}}$. Meanwhile, since SICE representation is obtained through optimisation, this may lead to computational or stability issues when working with high-dimensional features. With the current optimisation technique, it is recommended to handle features with the dimensions lower than 200. Certainly, this issue could be alleviated when advanced GLASSO is developed.

For the proposed kernel representation, without loss of generality, we conduct analysis with the most commonly used Gaussian RBF kernel. Given n d-dimensional feature vectors, $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, computing all the entries $\|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2$ $(i, j = 1, \cdots, d)$ has the complexity of $\mathcal{O}(nd^2)$, which is at the same level of computing a covariance matrix. Certainly, RBF kernel has an $\exp(\cdot)$ operation and needs a bit more time.

Both the resulting SICE and kernel representations have a fixed size of $d \times d$ independent of n, which is same as that of covariance matrix. Therefore, they will not incur extra computation in the subsequent operations like evaluating the similarity of these SPD-matrix-based representations or classifying them. Also, the two new representations retain the merit that they allow two sets of different sizes to be directly compared.

5.2 Convergence analysis

Above all, we recall an implicit assumption taken in covariance representation, that is, the visual feature vector $\boldsymbol{x}$ conforms to a distribution whose mean and variance exist. The following analysis is all based on this assumption.

By the law of large numbers, the empirical covariance will converge to the true covariance while the number of feature vectors tends to infinity with fixed feature dimension d (Park, 2007; Adamczak et al., 2010). This theoretically guarantees the stability and consistency of the obtained covariance representation. Observing this property naturally arises a question: does the SICE or kernel matrix representation have such a property? In other words, will the SICE/kernel matrix representation converge to certain "true SICE/kernel matrix" with the increase of feature vector number? The investigation into this issue is presented in the Appendices 3. In short, the availability of more feature vectors will make SICE/kernel representations more reliable and push them towards their true values. It has been shown in (Meinshausen and Bühlmann, 2006; Banerjee et al., 2008) that the SICE solution to Eq.(2) is guaranteed to converge to the global optimum for any λ. The RBF kernel representation will also converge with $n \to +\infty$.

5.3 Differences from existing work

As reviewed in Section 2, improving the effectiveness of covariance representation has been studied in the literature. Existing works in this aspect can roughly be categorised into three groups: i) improving the similarity evaluation of covariance representation with the theories of Riemannian manifold; ii) improving the quality of visual feature or image region to compute covariance representation; and iii) considering to model high-order statistics of features (Pang et al., 2008a),(Harandi et al., 2014b), (Cavazza et al., 2016), which is most related to our approach.

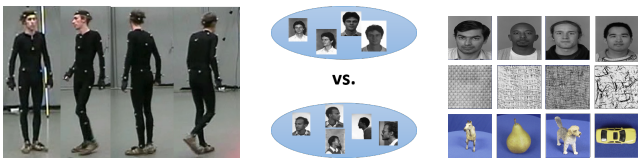
Our work is different from all the above three groups. Specifically, it is orthogonal to the works in the first and second groups, since the proposed new representations can work with any SPD-based similarity measure or visual feature set. Compared with the third group, we apply nonlinear kernel mapping to each *feature dimension* $\boldsymbol{f}_1, \cdots, \boldsymbol{f}_d$, instead of each *feature vector* $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$. The resulting representation maintains the same dimensions $(d \times d)$ as the original covariance representation, and does not need to design any special similarity measure. Also, it runs as efficiently as the original covariance representation. In addition, note that our work utilises kernel matrix to represent an individual feature set. This is different from existing works that develop kernel functions to measure the similarity between two feature sets (Póczos et al., 2012; Koniusz et al., 2016; Cavazza et al., 2017a, 2019).

The properties of the proposed representations and some of the competing ones are summarised in Table 1. As seen, the proposed SICE representation (SICE-RP in short) and kernel-based representation (Ker-RP in short) possess several desirable properties.

It is worth noting that dimensionality reduction is also a promising approach to the issue of small number of feature vectors and high dimensionality that we are trying to address in this paper. The new methods developed in our work are complementary to, rather than competing with, the approach of dimensionality reduction, i.e., the proposed methods can be readily applied to the features obtained by dimensionality reduction. In addition, addressing the issue of high dimensionality is not our sole motivation, and exploiting more advanced feature representations is an equally, if not more, important motivation of our work. The exploration of "structure sparsity" by SICE and the non-linearity in kernel based representation is also beneficial even when the issue of small number of feature vectors is not obvious.

Table 1 Summary of the differences between the proposed SPD representations and the competing SPD-based methods.

	Robust to small number of feature vectors & high dimensionality	Incorporated prior knowledge?	Not dependent on specific similarity measure	Guaranteed to be nonsingular and SPD	Able to model nonlinear feature relationship	Free of parameter tuning
Cov-RP (Tuzel et al., 2006)	×	×	√	×	×	√
Cov-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014b)	×	√	×	×	√	×
Cov3DJs Hierarchy (Hussein et al., 2013)	×	√	√	×	×	×
RSR-ML (Harandi et al., 2014a)	×	√	×	√	×	×
Square-Root-Cov-RP (Wang et al., 2019c)	√	√	√	×	×	√
Cov$^{\gamma}$ (Koniusz et al., 2013)	√	√	√	×	×	×
Kernelised-Cov (Cavazza et al., 2016)	√	√	√	×	√	×
SICE-RP (proposed)	√	√	√	√	×	×
Ker-RP (proposed)	√	√	√	√	√	×



(a) (b) (c)

Fig. 1 Illustration of three vision tasks investigated in this experiment. (a) shows an example of skeletal action sequences used in two tasks: human action recognition (images from HDM05 (Müller et al., 2009)); (b) illustrates image set classification. A set of images, rather than individual images, are classified as a whole (images from FERET (Phillips et al., 2000)); (c) plots three example object classification tasks (images from FERET (Phillips et al., 2000), Brodatz (Randen and Husoy, 1999), and ETH80 (Leibe and Schiele, 2003)).

Table 2 Summary of three skeletal human action recognition data sets.

Data set	#Feature dimensions (d)	#Frames per instance (n)
SBU Kinect (Yun et al., 2012a)	45	$10 \sim 46$
HDM05 (Müller et al., 2009)	93	$30 \sim 700$
NTU RGB+D (Shahroudy et al., 2016)	75	$10 \sim 300$

6 Experimental result

Regarding the proposed Ker-RP, the commonly used Gaussian RBF kernel is used for representation due to its flexibility in modeling high degree feature relationships (Hsu et al., 2003) and the resulting representation is denoted as Ker-RP-RBF. This is also in consistence with the fact that RBF kernel is usually recommended as the first choice in non-linear SVM classifier (Hsu et al., 2003). A convenient way is used to resolve Ker-RP-RBF normalisation and facilitate the setting of a uniform parameter β over all samples. That is, given a sample (say, an action sequence of n frames), all of its $\|\boldsymbol{f}_i - \boldsymbol{f}_j\|$ values $(i, j = 1, \cdots, n)$ will be divided by the average of the $\frac{n(n-1)}{2}$ pairwise Euclidean distances for computing the RBF kernel, where $\boldsymbol{f}_i$ denotes the i-th feature dimension. A nonlinear SVM classifier is used in all experiments except on ILSVRC2012 data set. The log-Euclidean kernel, a commonly used kernel function on SPD matrices, is employed for the SVM. The log-Euclidean kernel function is defined as $k(\boldsymbol{X}, \boldsymbol{Y}) = \exp\left(-\eta \| \log(\boldsymbol{X}) - \log(\boldsymbol{Y}) \|_F^2\right)$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two SPD matrices and $\log(\cdot)$ denotes the matrix logarithm. To ensure a fair comparison, all algorithmic parameters on all data sets, including the regularisation parameter in SVM, η in the log-Euclidean kernel for SVM, the parameters in the RBF kernels for generating the proposed kernel-based representation, and the sparsity parameter λ for the proposed SICE representation, are tuned by 5-fold cross-validation on the training set only.

The proposed SICE-RP and Ker-RP-RBF are compared with covariance representation (Cov-RP) and the

comparable methods on three types of computer vision tasks, as illustrated in Fig. 1. The first task is skeletal human action recognition, the second is image set classification and the third is object recognition. In addition to these three supervised learning tasks, the fourth task is an unsupervised retrieval of skeletal human actions reported in the Appendices 1.2.

6.1 Result on skeletal human action recognition

In the literature, covariance representation has been commonly evaluated with skeletal human action recognition. This is followed in our work. Three benchmark data sets, including HDM05 (Müller et al., 2009), SBU Kinect (Yun et al., 2012a) and NTU RGB+D (Shahroudy et al., 2016), are used in this experiment. For all of them, we only use the *skeleton* data while other data (e.g., depth maps or RGB videos) are not utilised. For the competing methods in comparison, the features used are explicitly listed. Information on these data sets is summarised in Table 2 and the details will be explained in each data set subsection. As seen, the number of frames per instance, n, could be smaller than $(d+1)$, which causes singularity when using Cov-RP. In this case, we follow the literature to append a small regulariser $\rho\boldsymbol{I}$ (e.g., $\rho = 10^{-7}$) to the obtained representation.

6.1.1 Result on HDM05 data set

HDM05 consists of around 1500 instances from over 100 motion classes. Most classes have 10 to 50 realisations of five actors named "bd", "bk", "dg", "mm" and "tr". We use two subjects "bd" and "mm" for training while the remaining three for test, and the 3D coordinates of each joint are used as the frame features by following (Harandi et al., 2014a). To compare with existing works, we conduct two experiments. Firstly, we use 14 classes of this data set, and report the result in the left column of Table 3. In addition to Cov-RP, the results of several other methods in the literature are also quoted, in which CDL (Wang et al., 2012), RSR (Harandi et al., 2012) and RSR-ML (Harandi et al., 2014a) use covariance-based representations as well. As seen from Table 3, Cov-RP shows quite competitive performance and outperforms three quoted methods. Cov-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014b) uses an infinite-dimensional covariance matrix in a kernel-induced feature space as representation. However, it does not perform well as Cov-RP, although it is better than several of the quoted methods. In contrast, Square-Root-Cov-RP (Wang et al., 2019c) and Cov$^{\gamma}$ (Koniusz et al., 2013) effectively improve the performance of Cov-RP. Our two proposed methods demonstrate remarkable performance, both SICE-RP and Ker-RP-RBF achieving a high classification accuracy of 96.8%, which is close to the state-of-the-art performance achieved by Kernelised-Cov (Cavazza et al., 2016), ϕ_p (Cavazza et al., 2019) and Log-Cov-Net (Cavazza et al., 2017b) on this data set. The superior performance of these methods could probably be attributed to the nonlinearity modeling by developing kernel approximation or multiple-layer networks especially for action data.

To further verify their effectiveness, we conduct a comparison on all the 112 action classes[2]. As shown in the right column of Table 3, although the significant increase on the total number of action classes reduces the overall classification accuracy, the two proposed methods still outperform most of the other ones in comparison. Specifically, SICE-RP and Ker-RP-RBF achieve significant improvements of 8.7 and 7.3 percentage points over Cov-RP, respectively, indicating the efficacy of exploiting the prior knowledge or modeling nonlinearity. Note that the result of Cov-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014b) is not reported for the all-class setting since it cannot be obtained in 35 hours on our computing facility.

Table 3 Comparison on HDM05 data set (Two experiments).

Methods in comparison	14 classes Accuracy	All classes Accuracy
CDL (Wang et al., 2012)	79.8	50.4†
RSR (Harandi et al., 2012)	76.1	-
RSR-ML (Harandi et al., 2014a)	81.9	40.0†
Kernelised-Cov (Cavazza et al., 2016)	98.1	-
CKA (Cavazza et al., 2017a)	-	65.0
ϕ_p (Cavazza et al., 2019)	99.1	72.0
Log-Cov-Net (Cavazza et al., 2017b)	99.1	72.0
Cov-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014b)	82.5	-
Cov-RP (Tuzel et al., 2006)	91.5	58.9
Square-Root-Cov-RP (Wang et al., 2019c)	92.4	65.1
Cov$^{\gamma}$ (Koniusz et al., 2013)	95.2	66.9
SICE-RP (proposed)	96.8	67.6
Ker-RP-RBF (proposed)	96.8	66.2

⋆The result of Cov-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014b) is not obtained in 35 hours.
†Obtained by this work with the code of (Wang et al., 2012) and (Harandi et al., 2014a).

[2] There are 130 classes in the original data set, among which 18 classes having less than five actors are removed.

6.1.2 Result on SBU Kinect data set

SBU Kinect data consists of eight types of two-person interactions performed by seven participants. There are two actors in each action sequence and 15 joint locations (x,y,z) for each actor, resulting in 90 features for each frame. Similar to (Yun et al., 2012a), the absolute distances between the corresponding joint locations of the two actors are used as features, i.e., [abs(x101 - x201), abs(y101 - y201), abs(z101 - z201), ..., abs(x115 - x215), abs(y115 - y215), abs(z115 - z215)], where x101 denotes the location x of the 01 joint for actor 1. In doing so, the final feature vector used to compute the proposed SICE- and kernel-based representations is of 45 dimensions. The 5-fold cross validation which is predefined in the data set (Yun et al., 2012a) is conducted. The comparison result is reported in Table 4. All methods in comparison are categorised into two groups with respect to whether the features are hand-crafted (i.e., no feature representation learning) or learned from data with deep learning techniques (i.e., with feature representation learning). Specifically, the upper part contains the methods that do not involve "feature representation learning". Note that the two proposed representations belong to this group. Moreover, these two representations do not use any temporal information of skeleton. The lower part of this table lists the methods that learn feature representation from data and they utilise the temporal information of skeleton. The following results can be observed.

1) Traditional covariance representation Cov-RP performs reasonably well (86.05%) and is comparable to some RNN based methods, e.g., Deep LSTM (Zhu et al., 2016) (86.0%). This demonstrates the effectiveness of SPD-matrix-based representation;

2) The proposed SICE- and kernel-based representations outperform Cov-RP by a large margin of 2.6 and 8.6 percentage points, respectively. This again shows their advantage over Cov-RP;

3) Particularly, the kernel-based representation achieves very competitive performance of 94.64%, even exceeding most of the RNN-based methods except Multilayer LSTM (Zhang et al., 2017b) and View-adaptive LSTM (Zhang et al., 2017a). Note that Multilayer LSTM (Zhang et al., 2017b) and View-adaptive LSTM (Zhang et al., 2017a) are two sophisticated models specially designed for skeletal action recognition. Specifically, Multilayer LSTM (Zhang et al., 2017b) builds a 3-layer LSTM model to exploit temporal relations between joints, while View-adaptive LSTM (Zhang et al., 2017a) designs a view-adaptive RNN model to capture the same action from different viewpoints. In contrast, the SICE- and kernel-based representations are proposed as generic visual descriptors. They can be generally applied not only to skeletal action recognition but also to image recognition and image set classification, as will be demonstrated in this work. These tasks are certainly not in the scope of the above RNN- or LSTM-based models.

Based on the above result, the effectiveness and advantage of the proposed SICE- and kernel-based representations with respect to Cov-RP can be validated on the SBU Kinect.

Table 4 Comparison of classification performance on SBU-Kinect data set.

Methods in comparison	Accuracy
No feature representation learning	
Yun et al. (Yun et al., 2012b)†	80.3
CHARM (Li et al., 2015)	83.9
Ji et al. (Ji et al., 2014)†	86.9
Cov-RP	86.05
SICE (proposed)	88.68
Ker-RP-RBF (proposed)	**94.64**
With feature representation learning	
HBRNN-L ((Du et al., 2015)†	80.35
Deep LSTM (Zhu et al., 2016)	86.0
Co-occurrence LSTM (Zhu et al., 2016)	90.41
STA-LSTM (Song et al., 2017)	91.5
ST-LSTM (Liu et al., 2016)	93.3
Clips + CNN + MTLN (Ke et al., 2017)	93.57
Context-aware LSTM (Liu et al., 2017)	94.1
View-adaptive LSTM (Zhang et al., 2017a)	97.2
Multilayer LSTM (Zhang et al., 2017b)	**99.02**

† denotes that the results are from (Zhang et al., 2017b).

6.1.3 Result on NTU RGB+D data set

NTU RGB+D (Shahroudy et al., 2016) data set is a large-scale multi-modality data set collected by Microsoft Kinect v2 sensors for human action recognition. There are over 56 thousand video samples collected from 40 distinct subjects. These video samples are categorized into 60 action classes, including daily, mutual, and health-related actions. On this data set, two training/test schemes are used in our evaluation by following the literature (Shahroudy et al., 2016). The first type is cross-subject evaluation. In this evaluation, the 40 subjects are split into two training and test groups with 20 subjects in each. The following subject IDs are assigned to the training group: 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38, and the rest are for test. The second type is cross-view evaluation. In this scheme, the samples recorded by camera 1 are used for test while those recorded by cameras 2 and 3 are used for training. Two kinds features are used. First, the 3D coordinates of each joint are used as the frame features by following (Shahroudy et al., 2016). The second feature type is the features extracted from Graph

Convolutional Networks trained on NTU RGB+D data set. Specifically, the Spatial Temporal Graph Convolutional Networks (ST-GCN[3]) from (Yan et al., 2018) and Two-Stream Adaptive Graph Convolutional Networks (2s-AGCN[4]) from (Shi et al., 2019) developed for skeleton-based action recognition are utilised. The output ($2 \times 256 \times 75 \times 25$ D) of the final convolutional layer before the fully connected layer is stacked into 3750 features with 256 D and used to compute Ker-RP-RBF (256×256 D) for classification. For the features from 2s-AGCN, the final classification is determined by combining the classification scores of joint and bone streams by following the literature 2s-AGCN (Shi et al., 2019). As seen in Table 5, similar to the result on SBU Kinect data set, the methods are categorised into upper part that does not involve feature representation learning and the lower part that learns feature representation from data and utilises the temporal information of skeleton. The following observations are obtained: 1) In the upper part of this table, SICE- and kernel-based representations achieve competitive performance and are comparable to the state-of-the-art methods. This again demonstrates their effectiveness. The promising performance of SCK-DCK (Koniusz et al., 2016) could probably be attributed to the speciality in their representation, which is specially designed for skeletal action data and the spatio-temporal information between joints and dynamics of sequences are explicitly modeled. In contrast, our proposed methods are proposed as general-purpose representation methods and their performance relies on the quality of features used. When a simple concatenation of joint locations is used as feature, the spatio-temporal information between joints and dynamics of sequences is not fully represented as in SCK-DCK (Koniusz et al., 2016). When such information is critical in differentiating actions in sophisticated data sets, the performance of our proposed method may be restricted by the feature while SCK-DCK (Koniusz et al., 2016) could obtain better performance in this case. 2) Compared with the methods in the lower part of this table, the proposed SICE- and kernel-based representations still outperform several RNN- and LSTM-based methods (e.g., LSTM (Shahroudy et al., 2016) and HBRNN-L (Du et al., 2015) shown with an underscore), even though they do not utilise any skeleton temporal information. 3) If advanced ST-GCN features are used, applying Cov-RP improves the performance of ST-GCN as expected, boosting the performance from 81.5% to 82.0% in cross-subject protocol and from 88.3% to 89.9% in cross-view protocol. The proposed Ker-RP-RBF further increases the performance to 82.9% in cross-subject protocol and 90.9% in cross-view protocol, obtaining improvements of 1.4 and 2.6 percentage points respectively over the ST-GCN. Similar conclusion can be drawn from the experiment with features from 2s-AGCN. As seen, Cov-RP boosts the performance of 2s-AGCN from 88.5% to 88.8% in cross-subject protocol and from 95.1% to 95.3% in cross-view protocol. Square-Root-Cov-RP (Wang et al., 2019c) and cov^{γ} (Koniusz et al., 2013) can even achieve slightly better results in cross-subject protocol and comparable results in cross-view protocol. The proposed Ker-RP-RBF further increases the performance to 89.2% in cross-subject protocol and 95.5% in cross-view protocol, obtaining the best performance among the methods in comparison. This shows that with advanced features, the proposed method is able to achieve the state-of-the-art performance.

Following the discussion on SBU Kinect experiment, we highlight that skeletal action recognition is used here as an example visual recognition task to show the effectiveness of SICE- and kernel-based representations, instead of competing them with the models like RNN that are specially designed for this kind of task. In other words, the SICE- and kernel-based representations are proposed as generic visual descriptors. As aforementioned, they can be generally applied not only to skeletal action recognition but also to other vision tasks that are not in the scope of the above RNN- or LSTM-based models. Considering the above points, the result on NTU RGB+D can serve the purpose of demonstrating the effectiveness of the two proposed representations.

In sum, as seen from the above experimental results on three different human action data sets, the proposed SICE-RP and Ker-RP not only improve Cov-RP significantly and consistently, but also achieve competitive performance on these data sets. This well verifies the advantage and necessity of considering prior knowledge and nonlinearity for SPD-based visual representations. In the Appendix 1, additional experiments are conducted on three more data sets, including MSR-Action3D, MSR-DailyActivity3D and MSRC-12, and similar results can be obtained.

6.2 Result on image set classification

An image set is a collection of images belonging to the same class but with variation, for example, images of the same object or facial images of the same person under different views. It is the image set, rather than an individual image therein, that will be classified. Covariance matrix has been used to model an image set (Wang et al., 2012). Now we compare it with the proposed Ker-RP, following the experimental settings

[3] `https://github.com/yysijie/st-gcn`
[4] `https://github.com/lshiwjx/2s-AGCN`

Table 5 Comparison of classification performance on NTU RGB+D data set.

Methods in comparison	Cross Subject	Cross View	Feature based on
No feature representation learning			
HOG^2 (Ohn-Bar and Trivedi, 2013)	32.2	22.3	Depth
Super Normal Vector (Yang and Tian, 2014)	31.8	13.6	Depth
HON4D (Oreifej and Liu, 2013)	30.6	7.3	Depth
Lie Group (Vemulapalli et al., 2014)	50.1	52.8	Skeleton
Skeletal Quads (Evangelidis et al., 2014)	38.6	41.4	Skeleton
Dynamic Skeletons (Hu et al., 2015)	60.2	65.2	Skeleton
ϕ_p (Cavazza et al., 2019)	60.9	63.4	Skeleton
Log-Cov-Net (Cavazza et al., 2017b)	60.9	63.4	Skeleton
Cov-RP (Tuzel et al., 2006)	61.8	63.8	Skeleton
Square-Root-Cov-RP (Wang et al., 2019c)	62.6	66.3	Skeleton
Cov^γ (Koniusz et al., 2013)	64.2	67.7	Skeleton
SICE-RP (proposed)	63.9	67.2	Skeleton
Ker-RP-RBF (proposed)	64.4	68.1	Skeleton
SCK-DCK (Koniusz et al., 2016)*	72.8	74.1	Skeleton
With feature representation learning			
HBRNN-L (Du et al., 2015)	59.1	64.0	Skeleton
1 Layer RNN (Shahroudy et al., 2016)	56.0	60.2	Skeleton
2 Layer RNN (Shahroudy et al., 2016)	56.3	64.1	Skeleton
1 Layer LSTM (Shahroudy et al., 2016)	59.1	66.8	Skeleton
2 Layer LSTM (Shahroudy et al., 2016)	60.7	67.3	Skeleton
1 Layer P-LSTM (Shahroudy et al., 2016)	62.1	69.4	Skeleton
2 Layer P-LSTM (Shahroudy et al., 2016)	62.9	70.3	Skeleton
DSSCA-SSLM (Shahroudy et al., 2017)	74.9	N.A.	RGB + Depth
ST-LSTM (Liu et al., 2016)	69.2	77.7	Skeleton
Multilayer LSTM (Zhang et al., 2017b)	70.3	82.4	Skeleton
Context-aware LSTM (Liu et al., 2017)	74.4	82.8	Skeleton
Temporal Sliding LSTM (Lee et al., 2017)	74.6	81.3	Skeleton
View-adaptive LSTM (Zhang et al., 2017a)	79.4	87.6	Skeleton
Clips + CNN + MTLN (Ke et al., 2017)	79.6	84.8	Skeleton
IndRNN (Li et al., 2018)*	83.0	89.0	Skeleton
ST-GCN (Yan et al., 2018)	81.5	88.3	Skeleton
ST-GCN + Cov-RP (Tuzel et al., 2006)	82.0	89.9	Skeleton
ST-GCN + Ker-RP-RBF (proposed)	82.9	90.9	Skeleton
Deep Bilinear (Hu et al., 2018)*	85.4	90.7	Skeleton
SR-TSL (Si et al., 2018)	84.8	92.4	Skeleton
2s-AGCN (Shi et al., 2019)	88.5	95.1	Skeleton
2s-AGCN + Cov-RP (Tuzel et al., 2006)	88.8	95.3	Skeleton
2s-AGCN + Square-Root-Cov-RP (Wang et al., 2019c)	88.9	95.3	Skeleton
2s-AGCN + cov^γ (Koniusz et al., 2013)	89.0	95.3	Skeleton
2s-AGCN + Ker-RP-RBF (proposed)	**89.2**	**95.5**	Skeleton

The *underline* denotes that these RNN- and LSTM-based methods are outperformed by the proposed SICE- and kernel-based representations.

* The result is quoted from (Wang et al., 2019a).

in (Wang et al., 2012). Three data sets are tested, including ETH80 (Leibe and Schiele, 2003), CMU MoBo (Gross and Shi, 2001), and YouTube Celebrities (Wolf et al., 2011). ETH80 has eight categories, with ten objects per category. For each object, there are 41 images corresponding to different views. CMU MoBo has 96 video sequences of 24 subjects, and YouTube Celebrities consists of 1910 video clips from 47 subjects, where face images of each subject are collected by face detectors. Images in all three data sets are resized to 20×20 and pixel intensities are used as features.

The training and test sets are created as follows. For CMU MoBo, all face images detected from the same video sequence form an image set. One image set is randomly selected from each subject for training, and the remaining image sets are for test. For YouTube, three image sets are randomly chosen from each subject for training, and another six sets are randomly chosen for test. In ETH80, the ten objects in a category are randomly halved into training and test sets. For each object, the 41 images of different views form an image set. The Ker-RP and Cov-RP are used to represent each image set. In total, 10 training and test pairs are created for each data set.

Following (Wang et al., 2012), we use Partial Least Squares (PLS) for classification and the code is down-

loaded from that work[5]. Table 6 reports the classification accuracy averaged on the 10 partitions for each method. Ker-RP-RBF achieves the best classification performance on ETH80, outperforming Cov-RP by 2 percentage points and Cov-$J_{\mathcal{H}}$-SVM by 2.3 percentage points. On CMU MoBo, it still well improves over Cov-RP and Cov-$J_{\mathcal{H}}$-SVM. On YouTube, Ker-RP-RBF achieves comparable performance to Cov-RP but clearly outperforms Cov-$J_{\mathcal{H}}$-SVM. Also, as shown in the last column, Ker-RP-RBF gives the overall best performance on the three data sets. Note that SICE-RP is not included in Table 6. This is because SICE-RP cannot be reliably obtained by the GLASSO optimisation process when feature dimensions exceed 400 and the number of feature vectors in an image set is much smaller, as in the case of the three data sets.

We are aware of that higher performance has been reported in (Hayat et al., 2017) on these three image set data sets. However, that work designs a classification strategy especially for image set classification task. In contrast, our work focuses on improving covariance representation with generic applications. Also, their settings, such as image size, are different from those in this work. Therefore, the work in (Hayat et al., 2017) is not included in the comparison to avoid confusion.

Table 6 Comparison on three data sets for image set classification.

Methods	ETH80	CMU MoBo	YouTube	Average
Cov-RP (CDL (Wang et al., 2012))	96.5	94.1	70.1	86.9
DMK (Sun et al., 2017)	96.8	–	–	–
Cov-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014b)	96.2	95.2	63.7	85.0
Ker-RP-RBF (proposed)	98.5	95.9	70.0	**88.1**

6.3 Result on object classification

6.3.1 With hand-crafted features

We have verified the effectiveness of the proposed representations on skeletal human action recognition. As shown, for that specific task, the number of feature vectors is relatively small while the feature dimensionality is high, and the prior knowledge (i.e., structure sparsity due to human kinematic configuration) is clear. In this experiment, we further investigate the proposed representations for the tasks where the feature dimensionality is usually lower while a larger number of feature vectors are available. In addition, there is no well-perceived prior knowledge as the previous case.

Three data sets are used, including Brodatz (Randen and Husoy, 1999) for texture classification, FERET (Phillips et al., 2000) for face recognition, and ETH80 (Leibe and Schiele, 2003) for object categorisation. These data sets are traditionally used in the literature to evaluate the object classification performance of covariance representation. Brodatz contains 112 textured images. Following (Harandi et al., 2012), each image is partitioned into 64 non-overlapping sub-images as one texture class, and these sub-images will be classified in the task. For FERET, we use the "b" subset of 200 subjects. Each has 10 images with various poses and illumination conditions. ETH80 was used for image set classification in Section 6.2, but here each individual image is viewed as a training or test sample and classified.

For all three data sets, every image/sub-image is scaled to a uniform size of 64×64 and a 43-dimensional feature vector is extracted around each pixel, including its intensity, x and y coordinates, and a set of Gabor features (8 orientations and 5 scales) by following (Harandi et al., 2012). Note that in this experiment, the covariance in Cov-RP is estimated from 4096 (64×64) feature vectors, which is sufficiently large compared with the feature dimensions 43. Therefore, Cov-RP will not encounter the problem of small number of feature vectors in this experiment. For each data set, it is randomly halved into training and test subsets. This is repeated 20 times to obtain average classification performance.

As seen in Table 7, Cov-RP demonstrates reasonably good performance. Meanwhile, SICE-RP still attains very competitive performance against Cov-RP on the three data sets, with an improvement of 3.5 percentage points on face recognition. This is consistent with the principle of "Bet on sparsity" (Hastie et al., 2005) and indicates that exploiting structure sparsity does not necessarily hurt the performance of representation, even when the feature dimensionality is low and the number of feature vectors is large. It can be a safe option for various applications. Ker-RP-RBF again achieves the highest accuracy and outperforms Cov-RP by 3.7 and 4.4 percentage points on Brodatz and FERET. This indicates the effectiveness of the proposed Ker-RP-RBF. As previous, Cov-$J_{\mathcal{H}}$-SVM is not included in the comparison, because it becomes time-consuming when the number of feature vectors, n, is large.

We notice that higher performance has been reported in the literature, e.g., 98.72% in ELBCM (Romero

[5] The work (Wang et al., 2012) also investigates Linear Discriminant Analysis. However, PLS always outperforms LDA as shown in that work.

et al., 2013), 97.9% in L2ECM (Li and Wang, 2012), 97.7% in Cov-RP (Tuzel et al., 2006) and 99.9% in TOSST (Koniusz and Cherian, 2016) on Brodatz data set. The discrepancy is due to different settings used in these works and ours, so the performance may not be directly comparable. A key difference is that the above four methods use 160×160 or 320×320 subimages as samples while we use 80×80 subimages as samples by following the literature (Harandi et al., 2012). Smaller crops as samples make our case much more challenging than those methods, so the reported performance is lower. This challenging setting better demonstrates the superior performance of the proposed methods.

Table 7 Comparison on object classification data sets.

Methods	Brodatz (texture)	FERET (face)	ETH80 (object)
Cov-RP (Tuzel et al., 2006)	81.2	81.0	94.0
SICE-RP (proposed)	82.1	84.5	94.2
Ker-RP-RBF (proposed)	**84.9**	**85.4**	**94.8**

⋆The result of Cov-$J_{\mathcal{H}}$-SVM (Harandi et al., 2014b) is not obtained in 35 hours.

6.3.2 With learned features from CNN

Convolutional neural network (CNN) has demonstrated promising performance and become a dominant technique in various areas recently. The convolutional feature maps produced by deep networks can be viewed as a set of deep local descriptors extracted from an image. They can also be used to compute the covariance- and kernel-based representations. Through this experiment, we will demonstrate that deep learning and the proposed representation complement with each other in obtaining better classification performance.

The common notations and evaluation protocol are detailed below. In the experiment, we apply the proposed kernel-based representation method to CNN features on Describable Texture Datasets (DTD) (Cimpoi et al., 2014), PASCAL07 (Everingham et al., 2010) and ILSVRC2012, respectively. DTD data set is a texture database consisting of 5,640 images annotated with 47 describable attributes as labels. Along the data set, 10 splits of the data are provided for training and test. The classification accuracies are averaged over the 10 splits for comparison. PASCAL07 data set consists of 9,963 images belonging to 20 categories, and the standard training/test sets are predefined for each category. The mean average precision (mAP) over 20 categories are used for comparison. ILSVRC2012 contains 1.2 million images in 1000 categories, and the split of training set and validation set are provided.

On the two data sets of DTD and PASCAL07 (the case for ILSVRC2012 will be introduced shortly), we extract the last convolutional layer (28×28, 512) of VGG-19 (Simonyan and Zisserman, 2014) pre-trained on ImageNet ILSVRC2012 data set. The features are stacked into 512×28^2 and used to compute a 512×512 covariance representation or a (RBF) kernel-based representation, which are denoted by VGG-19 + Cov-RP and VGG-19 + Ker-RP-RBF, respectively. The above two representations are compared with those obtained by the sum-pooling and max-pooling operations with the same features, which are denoted by VGG-19 + Sum Pooling and VGG-19 + Max Pooling, respectively. As previous, a log-Euclidean kernel SVM classifier is equally trained with each SPD representation to perform image classification. We also extract the last fully-connected layer of VGG-19 (i.e., a 4096-dimensional feature vector) to train an RBF kernel SVM classifier, and it is denoted by VGG-19 (4096D vector).

Result on DTD data set. Table 8 (in DTD column) shows the results on DTD data set. The upper portion of this table quotes the state-of-the-art results from the comparable methods, while the lower portion lists the results of the methods implemented by this work. Furthermore, the lower portion consists of three subsections. The first subsection lists the results obtained by competing methods with various networks. The second subsection lists the results obtained by Cov-RP upon networks while the third subsection shows the corresponding results of Ker-RP-RBF.

As seen from the lower portion, the method VGG-19, which uses the final fully-connected layer (FC7, 4096-dimensional) as feature, obtains an accuracy of 66%. This result is better than that obtained by using the 4096-dimensional features from DeCAF (Donahue et al., 2014) shown in the upper portion. Also, it even wins a combination of DeCAF feature and Improved Fisher Vector (IFV), denoted as DeCAF+IFV (Cimpoi et al., 2014) in the table. This demonstrates the powerfulness of VGG-19 network.

When the sum pooling scheme is applied to the last convolutional layer to obtain the 512-dimensional features (denoted by VGG-19+Sum Pooling), the performance is improved to 68.9%. When Cov-RP is applied to the same convolutional layer, the accuracy increases to 69.8%, demonstrating the effectiveness of the SPD-matrix-based representation on deep learning features. With the proposed Ker-RP-RBF, the performance is further boosted to 72.7%, which is higher than most of the quoted methods and is comparable to the state-of-the-art methods (Cimpoi et al., 2016; Lin et al., 2017) which use multi-scaled image resolutions or end-to-end learning. This well verifies the efficacy of the proposed

kernel-based representation when working with deep learning features.

Result on PASCAL07 data set. Experiment is further conducted on PASCAL07 with more recent networks and pooling methods. As aforementioned, the upper portion of Table 8 (in PASCAL07 column) lists the competing methods and quotes their results from the literature. The lower portion reports the methods implemented by this work in three subsections, including competing methods, Cov-RP upon various networks and the Ker-RP-RBF counterparts.

As seen, the proposed Ker-RP-RBF achieves the accuracy of 89.8% with features from VGG-19, outperforming all the VGG-19 based methods which use sum-pooling, max-pooling, FV (Fisher Vector)-pooling, FC (Fully-connected layer), or Cov-RP. Also, it is better than the VGG-19 with fine-tuning on PASCAL07 (89.3%, denoted by VGG-19 + FT). With a similar protocol, the features from the last convolutional layer (i.e., Conv5_2, 14×14, 512) in ResNet-101 are used to compute a 512×512 covariance representation or a kernel-based representation for SVM classification, and they are denoted by ResNet-101 + Cov-RP and ResNet-101 + Ker-RP-RBF, respectively. These two representations are compared with ResNet-101. Similarly, we also extract the last layer of ResNet-101 (i.e., a 2048-dimensional feature vector) to train an SVM classifier and it is denoted by ResNet-101 (2048D vector). ResNet-101 + Cov-RP achieves 90.05%, and outperforms ResNet-101 (2048D vector, 87.2%) by a margin of 2.85 percentage points. ResNet-101 + Ker-RP-RBF further boosts the accuracy to 91.03%, which is higher than that of ResNet-101 by 3.83 percentage points. This result shows the effectiveness of the kernel-based representation when working with the convolutional features learned by modern CNN networks.

The above observation is further confirmed by extracting features from the ResNet-101 fine-tuned on PASCAL07, which is denoted by ResNet-101-FT. As seen, ResNet-101-FT (2048D vector) performs better than ResNet-101 (2048D vector) due to the fine-tuning process and is comparable to ResNet-101-FT, which directly performs classification with the fine-tuned ResNet-101 network. ResNet-101-FT + Cov-RP further improves ResNet-101-FT (2048D vector) and obtains 92.8%. As for ResNet-101-FT + Ker-RP-RBF, it achieves a performance of 93.7% and outperforms all the quoted methods in the upper portion of this table.

The proposed kernel-based representation is further evaluated with WILDCAT (Durand et al., 2017) network, which applies an advanced class-wise and spatial pooling scheme and achieves the state-of-the-art performance on PASCAL07. The features from the class-wise pooling layer (14×14, 160) are extracted from WILDCAT (Durand et al., 2017) to implement WILDCAT + Cov-RP and WILDCAT + Ker-RP-RBF. As seen, WILDCAT + Ker-RP-RBF still can improve over WILDCAT (Durand et al., 2017) and obtains the highest accuracy among all methods in comparison.

Result on ILSVRC2012 data set. Finally, the evaluation on ILSVRC2012 (Russakovsky et al., 2015) is also conducted. In this experiment, the publicly available pre-trained ResNet-101 model provided by VLFeat MatConvNet (Vedaldi and Lenc, 2015) is used as the baseline. With the same model, the last convolutional layer (conv5_2, 14×14, 512) is extracted to compute a 512×512 (RBF) kernel-based representation. This representation is then vectorised and used to train a single fully connected layer as a classifier, which is in the same form of the last classification layer in ResNet-101. In order to make a fair and easily reproducible evaluation, we use a single crop of image and a single model (i.e., the above publicly available pre-trained ResNet-101 model) in both training and test stages.[6]

The proposed kernel-based representation attains the top-1 accuracy of 76.81%, which is higher than that of ResNet-101 (76.36%) by 0.45 percentage point. Also, for the top-5 accuracy, the proposed kernel-based representation achieves 93.13%, and it is again higher than that of ResNet-101 (92.87%). Considering the level of the difficulty of ILSVRC2012, these improvements are valuable and indicative. They again demonstrate that applying the proposed kernel-based representation to the convolutional layer features of CNN models could further improve the performance of image classification.

In sum, with more advanced deep features, the proposed kernel-based representation is able to consistently outperform the covariance representation and achieve the state-of-the-art recognition performance. This again indicates the effectiveness of this representation.

Visualisation of the proposed representations. Before ending this experimental study, to provide an intuitive understanding of the proposed SICE-RP and Ker-RP in comparison with Cov-RP, Fig. 2 visualises the three representations of "Shoot with a pistol" action in MSRC-12 data set with explanation in the annotation. The configuration and names of joints are provided in the Appendix 7. Two more example actions, "Navigate to next menu" and "Kick to attack an enemy", are also visualised in the Appendix. From this visualisation, we can see that i) in comparison with Cov-RP, the proposed SICE-RP only shows a few direct and significant correlations, indicating that SICE-RP has a func-

[6] We notice that a higher baseline could be achieved by using an ensemble of multiple ResNet-101 models and multiple crops of images.

Table 8 Comparison of classification performance on DTD and PASCAL07 data sets.

Methods in comparison	DTD (Accuracy)	PASCAL07 (mAP)
Quoted methods		
DeCAF (Donahue et al., 2014)	52.5 ± 1.3	-
DeCAF + IFV (Cimpoi et al., 2014)	64.7 ± 1.7	-
Return of devil (Chatfield et al., 2014)	-	82.4
CNN:STOM (Wei et al., 2014)	-	85.2
VGG-19 + FC (Cimpoi et al., 2016)	65.3 ± 1.5	84.6
VGG-19 + FV-Multi-scale (Cimpoi et al., 2016)	$72.3 \pm 1.0^{\dagger}$	88.6
VGG-16 + FV-Multi-scale (Cimpoi et al., 2016)	$73.6 \pm 1.0^{\dagger}$	-
VGG-19 + FT (Simonyan and Zisserman, 2014)	-	89.3
VGG16 (Simonyan and Zisserman, 2014)	-	89.3
WELDON (Oquab et al., 2015)	-	90.2
SPLeaP (Kulkarni et al., 2016)	-	88.0
RRSVM (Wei and Hoai, 2016)	-	92.9
ResNet-101 (He et al., 2016)	-	89.8
B-CNN (Lin et al., 2017)	$72.9 \pm 0.8^{\S}$	-
Implemented by this paper		
Competing CNN methods		
VGG-19 (4096D vector)	66.0 ± 0.9	84.8
VGG-19 + Sum Pooling	68.9 ± 1.0	82.1
VGG-19 + Max Pooling	66.2 ± 1.2	87.0
ResNet-101 (2048D vector)	-	87.2
ResNet-101-FT	-	89.7
ResNet-101-FT (2048D vector)	-	90.0
WILDCAT (Durand et al., 2017)	-	94.04
CNN + Cov-RP		
VGG-19 + Cov-RP (Tuzel et al., 2006)	69.8 ± 0.7	86.8
ResNet-101 + Cov-RP (Tuzel et al., 2006)	-	90.05
ResNet-101-FT + Cov-RP (Tuzel et al., 2006)	-	92.8
WILDCAT (Durand et al., 2017) + Cov-RP (Tuzel et al., 2006)	-	94.10
CNN + Ker-RP-RBF		
VGG-19 + Ker-RP-RBF (proposed)	72.7 ± 1.0	89.8*
ResNet-101 + Ker-RP-RBF (proposed)	-	91.03*
ResNet-101-FT + Ker-RP-RBF (proposed)	-	93.7*
WILDCAT (Durand et al., 2017)+ Ker-RP-RBF (proposed)	-	**94.16***

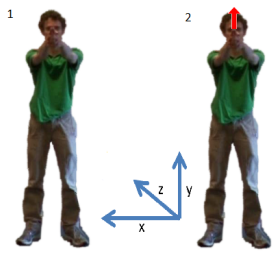
The * indicates the best performance in each comparable category with the same settings.
The † indicates that multi-scaled resolutions are used.
The § indicates that B-CNN is an end-to-end learning method.

tion of "focusing more on the essential and dominant correlations" rather than all the apparent correlations in Cov-RP. ii) the proposed Ker-RP-RBF shows much denser patterns than Cov-RP and SICE-RP because it actively extracts nonlinear relationship from each pair of feature dimensions to improve the expressiveness of the representation.

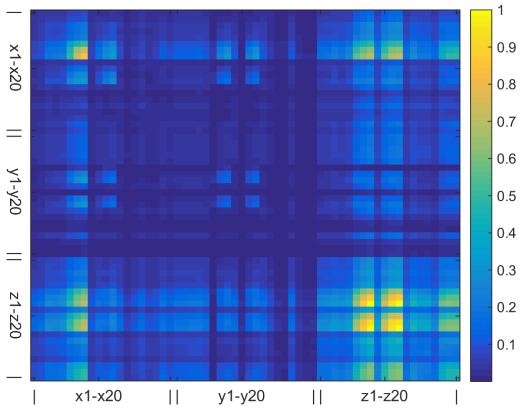
7 Conclusion and future work

In order to address the new issues encountered by covariance representation in visual recognition tasks, we analyse the essence of this representation and propose two new SPD-based visual feature representations. The SICE-based representation exploits prior knowledge to accurately model feature relationship, and the kernel-based representation characterises complicated nonlinear relationships among features. By discussing the properties of the proposed representations, their merits with respect to the state-of-the-art counterparts are manifested. And these merits are further experimentally verified via multiple visual recognition tasks. By moving beyond traditional covariance-based representation to pursue more advanced SPD-based representations, this work consistently shows that the proposed visual representations can attain higher recognition performance. In addition, the proposed representations are visualised to facilitate intuitive understanding.

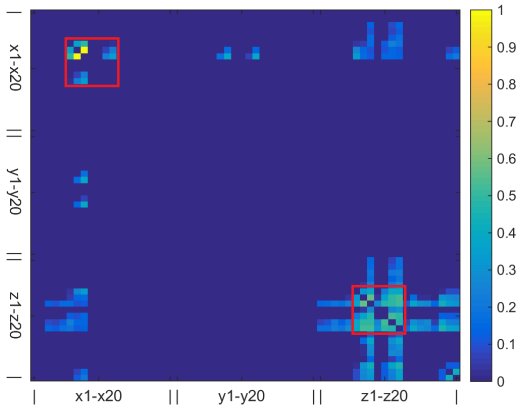
Several open issues are worth exploring along this line of research. Firstly, for the kernel-based representation, how to automatically choose, design or even learn the most appropriate kernel is an important issue to address, although the RBF kernel seems to be a good option in default. Secondly, in this work, structure sparsity is focused to exemplify the advantage of exploiting prior knowledge for feature representation. Other forms of knowledge shall be explored in further to demonstrate the power of this approach. Thirdly, the vari-



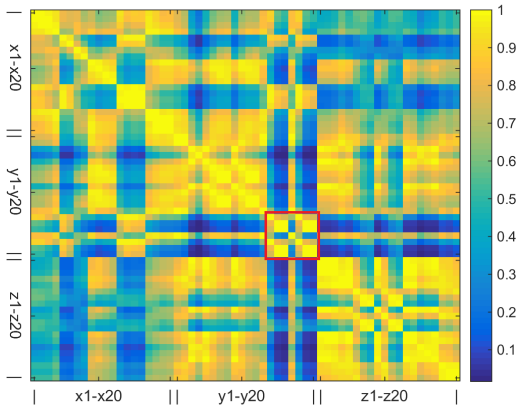
(a) **Shoot with a pistol** (Image from MSRC-12)



(b) Covariance Representation



(c) Proposed SICE-RP



(d) Proposed Ker-RP-RBF

Fig. 2 Visualisation of Cov-RP, SICE-RP and Ker-RP-RBF for a "Shoot with a pistol" action in MSRC-12 data set. The labels $x1-x20$, $y1-y20$, and $z1-z20$ denote the x, y, z-coordinates of the 20 skeletal joints, respectively. As seen, Cov-RP in (b) presents dense pairwise correlations between most joints. In contrast, the proposed SICE-RP in (c) only shows a few direct and significant correlations. This indicates that SICE-RP has a function of "focusing more on the essential and dominant correlations" rather than all the apparent correlations in Cov-RP. For example, the blocks in the two red bounding boxes in (c) indicate the interaction of the x-coordinates of two arms and the interaction of the z-coordinates of two arms, respectively. This is consistent with the "Shoot with a pistol" action, in which two hands mainly move along the x-axis and z-axis to hold together to form a pistol. The proposed Ker-RP-RBF in (d) shows much denser patterns than Cov-RP and SICE-RP because it actively extracts nonlinear relationship from each pair of feature dimensions to improve the expressiveness of the representation. For example, the red box in Ker-RP-RBF illustrates the nonlinear relationships between the right and left legs when they are alternately lifted up. Cov-RP fails to capture this pattern.

ous feature representations involved in this work model data from different perspectives and at different levels, and they could complement each other. How to adaptively fuse these representations for a given task becomes an interesting topic. For example, the nonlinearity in kernel-based representation and the sparsity in SICE representation could be integrated to further boost the representation capability. Last but not least, the effectiveness of the proposed representations will be further explored for more visual tasks and end-to-end learnable models in our future work.

8 Acknowledgement

The authors would like to thank Chang Tang for helping with preparing the action data sets and Ruiping Wang for sharing their processed image set data sets used in the experiments. Thanks also go to anonymous reviewers and editors for the constructive comments and the guidance in the revision of this paper. Lei Wang is the recipient of an Australian Research Council Discovery Project (project number DP200101289) funded by the Australian Government.

References

Adamczak R, Litvak A, Pajor A, Tomczak-Jaegermann N (2010) Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. Journal of the American Mathematical Society 23(2):535–561

Ali S, Basharat A, Shah M (2007) Chaotic invariants for human action recognition. In: IEEE International Conference on Computer Vision, IEEE, pp 1–8

Arsigny V, Fillard P, Pennec X, Ayache N (2006) Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic Resonance in Medicine 56(2):411–421

Banerjee O, Ghaoui LE, d'Aspremont A (2008) Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine Learning Research 9(Mar):485–516

Basser PJ, Mattiello J, LeBihan D (1994) Estimation of the effective self-diffusion tensor from the nmr spin echo. Journal of Magnetic Resonance, Series B 103(3):247–254

Cavazza J, Zunino A, San Biagio M, Murino V (2016) Kernelized covariance for action recognition. In: International Conference on Pattern Recognition, IEEE, pp 408–413

Cavazza J, Morerio P, Murino V (2017a) A compact kernel approximation for 3D action recognition. In: International Conference on Image Analysis and Processing, Springer, pp 211–222

Cavazza J, Morerio P, Murino V (2017b) When kernel methods meet feature learning: Log-covariance network for action recognition from skeletal data. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 33–40

Cavazza J, Morerio P, Murino V (2019) Scalable and compact 3D action recognition with approximated rbf kernel machines. Pattern Recognition 93:25–35

Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:14053531

Cimpoi M, Maji S, Kokkinos I, Mohamed S, , Vedaldi A (2014) Describing textures in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 3606–3613

Cimpoi M, Maji S, Kokkinos I, Vedaldi A (2016) Deep filter banks for texture recognition, description, and segmentation. International Journal of Computer Vision 118(1):65–94, DOI 10.1007/s11263-015-0872-3

Cirujeda P, Binefa X (2014) 4DCov: a nested covariance descriptor of spatio-temporal features for gesture recognition in depth sequences. In: International Conference on 3D Vision, IEEE, vol 1, pp 657–664

Cui Y, Zhou F, Wang J, Liu X, Lin Y, Belongie S (2017) Kernel pooling for convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2921–2930

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) Decaf: A deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning, pp 647–655

Dryden IL, Koloydenko A, Zhou D (2009) Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. The Annals of Applied Statistics pp 1102–1123

Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1110–1118

Durand T, Mordan T, Thome N, Cord M (2017) Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 642–651

Evangelidis G, Singh G, Horaud R (2014) Skeletal quads: Human action recognition using joint quadruples. In: International Conference on Pattern Recognition, IEEE, pp 4513–4518

Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88(2):303–338

Fasshauer GE (2011) Positive definite kernels: past, present and future. Dolomites Research Notes on Approximation 4(Special Issue on Kernel Functions and Meshless Methods):21–63

Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1933–1941

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441

Gao Y, Beijbom O, Zhang N, Darrell T (2016) Compact bilinear pooling. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 317–326

Gross R, Shi J (2001) The cmu motion of body (mobo) database. Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA

Guo K, Ishwar P, Konrad J (2010) Action recognition using sparse representation on covariance manifolds of optical flow. In: IEEE international conference on advanced video and signal based surveillance, IEEE, pp 188–195

Harandi MT, Sanderson C, Hartley R, Lovell BC (2012) Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In: European Conference on Computer Vision, Springer, pp 216–229

Harandi MT, Salzmann M, Hartley R (2014a) From manifold to manifold: geometry-aware dimensionality reduction for SPD matrices. In: European Conference on Computer Vision, Springer, pp 17–32

Harandi MT, Salzmann M, Porikli FM (2014b) Bregman divergences for infinite dimensional covariance matrices. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1003–1010

Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27(2):83–85

Hayat M, Khan SH, Bennamoun M (2017) Empowering simple binary classifiers for image set based face recognition. International Journal of Computer Vision pp 1–20

Haykin S (1998) Neural Networks: A Comprehensive Foundation (2nd Edition). Prentice Hall PTR., Upper Saddle River, NJ, USA

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 770–778

Hsu CW, Chang CC, Lin CJ, et al. (2003) A practical guide to support vector classification

Hu JF, Zheng WS, Lai J, Zhang J (2015) Jointly learning heterogeneous features for rgb-d activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 5344–5352

Hu JF, Zheng WS, Pan J, Lai J, Zhang J (2018) Deep bilinear learning for rgb-d action recognition. In: European Conference on Computer Vision, Springer, pp 335–351

Huang J, Zhang T, Metaxas D (2011) Learning with structured sparsity. The Journal of Machine Learning Research 12:3371–3412

Huang S, Li J, Sun L, Ye J, Fleisher A, Wu T, Chen K, Reiman E (2010) Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. NeuroImage 50(3):935–949

Hussein ME, Torki M, Gowayyed MA, El-Saban M (2013) Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In: International Joint Conference on Artificial Intelligence, pp 2466–2472

Ionescu C, Vantzos O, Sminchisescu C (2015) Matrix backpropagation for deep networks with structured layers. In: IEEE International Conference on Computer Vision, IEEE, pp 2965–2973

Jayasumana S, Hartley R, Salzmann M, Li H, Harandi M (2013) Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 73–80

Ji Y, Ye G, Cheng H (2014) Interactive body part contrast mining for human interaction recognition. In: IEEE International Conference on Multimedia and Expo Workshops, IEEE, pp 1–6

Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3D action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 4570–4579

Koller D, Friedman N (2009) Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning. The MIT Press

Koniusz P, Cherian A (2016) Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 5395–5403

Koniusz P, Yan F, Gosselin PH, Mikolajczyk K (2013) Higher-order occurrence pooling on mid-and low-level features: Visual concept detection. HAL-Inria

Koniusz P, Cherian A, Porikli F (2016) Tensor representations via kernel linearization for action recognition from 3D skeletons. In: European Conference on Computer Vision, Springer, pp 37–53

Kulkarni P, Jurie F, Zepeda J, Pérez P, Chevallier L (2016) Spleap: Soft pooling of learned parts for image classification. In: European Conference on Computer Vision, Springer, pp 329–345

Lee I, Kim D, Kang S, Lee S (2017) Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In: IEEE International Conference on Computer Vision, IEEE, pp 1012–1020

Lehrmann AM, Gehler PV, Nowozin S (2013) A non-parametric bayesian network prior of human pose. In: IEEE International Conference on Computer Vision, IEEE, pp 1281–1288

Leibe B, Schiele B (2003) Analyzing appearance and contour based methods for object categorization. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, vol 2, pp II–409

Li P, Wang Q (2012) Local log-euclidean covariance matrix (l2ecm) for image representation and its applications. In: European conference on computer vision, Springer, pp 469–482

Li P, Xie J, Wang Q, Zuo W (2017) Is second-order information helpful for large-scale visual recognition? In: IEEE International Conference on Computer Vi-

sion, IEEE, pp 2070–2078
Li S, Li W, Cook C, Zhu C, Gao Y (2018) Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 5457–5466
Li W, Wen L, Choo Chuah M, Lyu S (2015) Category-blind human action recognition: A practical recognition system. In: IEEE International Conference on Computer Vision, IEEE, pp 4444–4452
Lin TY, RoyChowdhury A, Maji S (2015) Bilinear cnn models for fine-grained visual recognition. In: IEEE International Conference on Computer Vision, IEEE, pp 1449–1457
Lin TY, RoyChowdhury A, Maji S (2017) Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(6):1309–1322
Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal lstm with trust gates for 3D human action recognition. In: European Conference on Computer Vision, Springer, pp 816–833
Liu J, Wang G, Hu P, Duan LY, Kot AC (2017) Global context-aware attention lstm networks for 3D action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, vol 7, p 43
Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. The annals of statistics pp 1436–1462
Müller M, Baak A, Seidel HP (2009) Efficient and robust annotation of motion capture data. In: ACM SIGGRAPH/Eurographics Symposium on Computer Animation, New Orleans, USA, pp 17–26
Ohn-Bar E, Trivedi M (2013) Joint angles similarities and hog2 for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 465–470
Oquab M, Bottou L, Laptev I, Sivic J (2015) Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 685–694
Oreifej O, Liu Z (2013) Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 716–723
Pang Y, Yuan Y, Li X (2008a) Effective feature extraction in high-dimensional space. IEEE Transactions on Systems, Man, and Cybernetics Part B, Cybernetics 38(6):1652–1656
Pang Y, Yuan Y, Li X (2008b) Gabor-based region covariance matrices for face recognition. IEEE Transactions on Circuits and Systems for Video Technology 18(7):989–993
Park J (2007) Digital Correlation Matrix in Multivariate Statistics and Its Application for Component Selection and Dynamic Correlation Modeling. ProQuest
Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10):1090–1104
Póczos B, Xiong L, Sutherland DJ, Schneider JG (2012) Nonparametric kernel estimators for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2989–2996
Porikli F, Tuzel O, Meer P (2006) Covariance tracking using model update based on lie algebra. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 728–735
Quang MH, San Biagio M, Murino V (2014) Log-hilbert-schmidt metric between positive definite operators on hilbert spaces. In: Conference on Neural Information Processing Systems, pp 388–396
Randen T, Husoy JH (1999) Filtering for texture classification: A comparative study. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(4):291–310
Romero A, Gouiffès M, Lacassagne L (2013) Enhanced local binary covariance matrices (elbcm) for texture analysis and object tracking. In: International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, pp 1–8
Russakovsky O, Deng J, et al (2015) ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3):211–252, DOI 10.1007/s11263-015-0816-y
Schölkopf B, Smola AJ, Bach F, et al. (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press
Shahroudy A, Liu J, Ng TT, Wang G (2016) NTU RGB+D: A large scale dataset for 3D human activity analysis. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1010–1019
Shahroudy A, Ng TT, Gong Y, Wang G (2017) Deep multimodal feature analysis for action recognition in rgb+ d videos. IEEE Transactions on Pattern Analysis and Machine Intelligence
Shi L, Zhang Y, Cheng J, Lu H (2019) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 12026–12035
Si C, Jing Y, Wang W, Wang L, Tan T (2018) Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: European Conference on Computer Vision, pp 103–118

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556

Smith SM, Miller KL, Salimi-Khorshidi G, Webster M, Beckmann CF, Nichols TE, Ramsey JD, Woolrich MW (2011) Network modelling methods for FMRI. Neuroimage 54(2):875–891

Song S, Lan C, Xing J, Zeng W, Liu J (2017) An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI Conference on Artificial Intelligence, pp 4263–4270

Sra S (2011) Positive definite matrices and the symmetric stein divergence. arXiv preprint arXiv:11101773

Sun H, Zhen X, Zheng Y, Yang G, Yin Y, Li S (2017) Learning deep match kernels for image-set classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3307–3316

Tabia H, Laga H, Picard D, Gosselin PH (2014) Covariance descriptors for 3D shape matching and retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 4185–4192

Tuzel O, Porikli F, Meer P (2006) Region covariance: A fast descriptor for detection and classification. In: European Conference on Computer Vision, Springer, pp 589–600

Tuzel O, Porikli F, Meer P (2008) Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(10):1713–1727

Vedaldi A, Lenc K (2015) Matconvnet – convolutional neural networks for matlab. In: ACM Int. Conf. on Multimedia

Vedaldi A, Zisserman A (2012) Efficient additive kernels via explicit feature maps. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3):480–492

Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3D skeletons as points in a lie group. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 588–595

Wang L, Zhang J, Zhou L, Tang C, Li W (2015a) Beyond covariance: Feature representation with nonlinear kernel matrices. In: IEEE International Conference on Computer Vision, IEEE, pp 4570–4578

Wang L, Huynh DQ, Koniusz P (2019a) A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing 29:15–28

Wang Q, Li P, Hu Q, Zhu P, Zuo W (2019b) Deep global generalized gaussian networks. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 5080–5088

Wang Q, Xie J, Zuo W, Zhang L, Li P (2019c) Deep cnns meet global covariance pooling: Better representation and generalization. arXiv preprint arXiv:190406836

Wang R, Guo H, Davis LS (2012) Covariance discriminative learning: A natural and efficient approach to image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2496–2503

Wang W, Wang R, Huang Z, Shan S, Chen X (2015b) Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2048–2057

Wei Y, Xia W, Huang J, Ni B, Dong J, Zhao Y, Yan S (2014) Cnn: Single-label to multi-label. arXiv preprint arXiv:14065726

Wei Z, Hoai M (2016) Region ranking svm for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2987–2996

Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 529–534

Wu Y, Ma B, Jia Y (2015) Differential tracking with a kernel-based region covariance descriptor. Pattern Anal Appl 18(1):45–59, DOI 10.1007/s10044-014-0430-6

Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI Conference on Artificial Intelligence

Yang X, Tian Y (2014) Super normal vector for activity recognition using depth sequences. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 804–811

Yuan C, Hu W, Li X, Maybank SJ, Luo G (2009) Human action recognition under log-euclidean riemannian metric. In: Asian Conference on Computer Vision, pp 343–353

Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012a) Two-person interaction detection using body-pose features and multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 28–35

Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012b) Two-person interaction detection using body-pose features and multiple instance learning. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 28–35

Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N (2017a) View adaptive recurrent neural networks

for high performance human action recognition from skeleton data. In: IEEE International Conference on Computer Vision, IEEE, pp 2136–2145

Zhang S, Liu X, Xiao J (2017b) On geometric features for skeleton-based action recognition using multilayer lstm networks. In: IEEE Winter Conference on Applications of Computer Vision, IEEE, pp 148–157

Zhu W, Lan C, Xing J, Zeng W, Li Y, Shen L, Xie X (2016) Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: AAAI Conference on Artificial Intelligence, pp 3697–3703

Zunino A, Cavazza J, Murino V (2017) Revisiting human action recognition: Personalization vs. generalization. In: International Conference on Image Analysis and Processing, Springer, pp 469–480